Investigating backtranslation for the improvement of English-Irish machine translation

Meghan Dowling, Teresa Lynn, & Andy Way

ADAPT Centre, School of Computing, Dublin City University

meghan.dowling@adaptcentre.ie, *teresa.lynn@adaptcentre.ie*, & *andy.way@adaptcentre.ie*

Abstract

In this paper, we discuss the difficulties of building reliable machine translation (MT) systems for the English-Irish (EN-GA) language pair. In the context of limited datasets, we report on assessing the use of backtranslation as a method for creating artificial EN-GA data to increase training data for use in state-of-the-art data-driven translation systems. We compare our results to our earlier work on EN-GA machine translation (Dowling et al. 2016; 2017; 2018) showing that while our own systems underperform with respect to traditionally reported automatic evaluation metrics, we provide a linguistic analysis to suggest that future work with domain-specific data may prove more successful.

***Keywords**: Machine translation; Gaeilge; English; automatic evaluation metrics*

1. Introduction

MT is a well-established tool in the post-editing environment of a professional translator. However, as a lesser-resourced and minority language, the Irish language has not enjoyed the benefits of technological advancements in the field of MT to the same extent that well-resourced languages (such as English) have. The status of Irish as the national and first official language of the Republic of Ireland, as well as an official European Union (EU) language, means a government requirement for all official documents and public services to be made accessible in both Irish and

English.[1] At present, the demand for bilingual content exceeds the productivity capabilities of translation services in Irish government departments and cannot be fully met by human translators alone. This will become more severe once the derogation granted to the Irish language runs out in 2021. Accordingly, we contend that this increasing imbalance between supply and demand necessitates a technology-orientated solution, which we explore in this paper. If done properly, EN-GA MT will be invaluable in meeting the language rights needs of Irish speakers. Our work aims to improve EN-GA MT so that it may be used as a more reliable practical aid in the production of translations at a national and European level. Our research focuses on examining a number of possible ways to improve EN-GA MT. This includes both experimenting with MT infrastructures and also investigating the current resources available for this language pair.

Both commonly used MT paradigms - statistical machine translation (SMT) and neural machine translation (NMT) - need to be 'trained on' a huge amount of data (i.e. high quality bilingual corpora) to produce quality translations. In recent years, there has been a shift towards the use of NMT, which is the most widely used MT paradigm at the moment. However, because NMT requires even more parallel (bilingual) data than SMT it exacerbates the problem of data sparsity (see Section 2 for a detailed explanation). Recently Dowling et al. (2018) carried out an assessment of the suitability of NMT with the EN-GA language pair. The preliminary results of this study showed that, although NMT results were not seen to rival that of SMT, improvements could be seen when steps were taken to optimise the NMT system.

The aim of the research presented in this paper is to build on this study by addressing the area of data sparsity through backtranslation, a method of creating artificial parallel data through the translation of monolingual data using pre-built MT systems (Poncelas et al. 2018 - see Section 6 for a detailed description). The premise of this method is that even if the data is not of human quality, the MT system can still draw benefits from the extra data. Another possible advantage

1 The Official Languages Act (2003) requires all official public information and services to be available in both Irish and English: http://www.irishstatutebook.ie/eli/2003/act/32/enacted/en/html

of backtranslation is that it can be used to combine SMT and NMT, and in theory combine the benefits of each approach.

SMT is a method of creating an MT model based on statistics and probability. SMT uses monolingual text to build a language model, which models what accurate sentences in the target language should look like. It also requires a large amount of bilingual text, known as a parallel corpus, to statistically compute the probability of translations. This is known as a translation model. See Hearne and Way (2011) for a detailed guide to SMT.

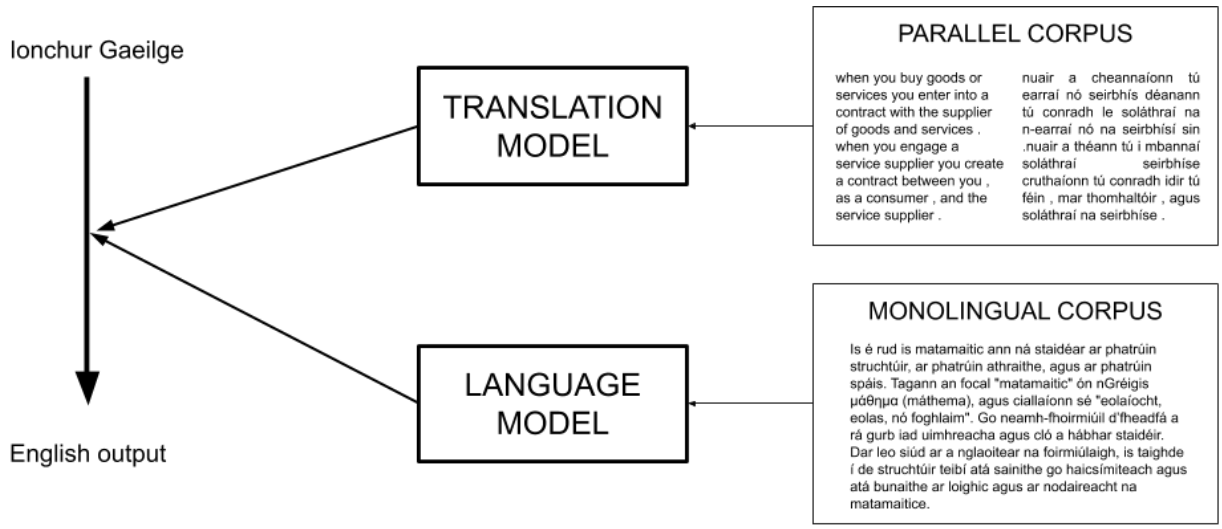

Figure 1: Simplified diagram of SMT

NMT (Sutskever et al. 2014; Bahdanau et al. 2015) also uses parallel text to train a translation model, but the model is trained using neural networks. There are a number of ways to train an NMT system, the most common following the 'encoder-decoder' methodology. A simplified diagram of an encoder-decoder NMT system can be seen in Figure 2. The input text is first *encoded* into a non-word representation suitable for translation – generally a vector of real numbers. This representation can then be *decoded* into the target-language text (i.e. translated text). See Forcada (2017) for a detailed introduction to NMT. Some reported strengths of NMT include a perceived increase in fluency and a higher accuracy according to automatic metrics over a variety of language pairs (Castilho et al 2017; Bojar et al. 2016). Some weaknesses of NMT are

a loss of semantics (the output looks fluent but has a different meaning - see Section 7.1) and overtranslation (the same word appearing more than once in the output).

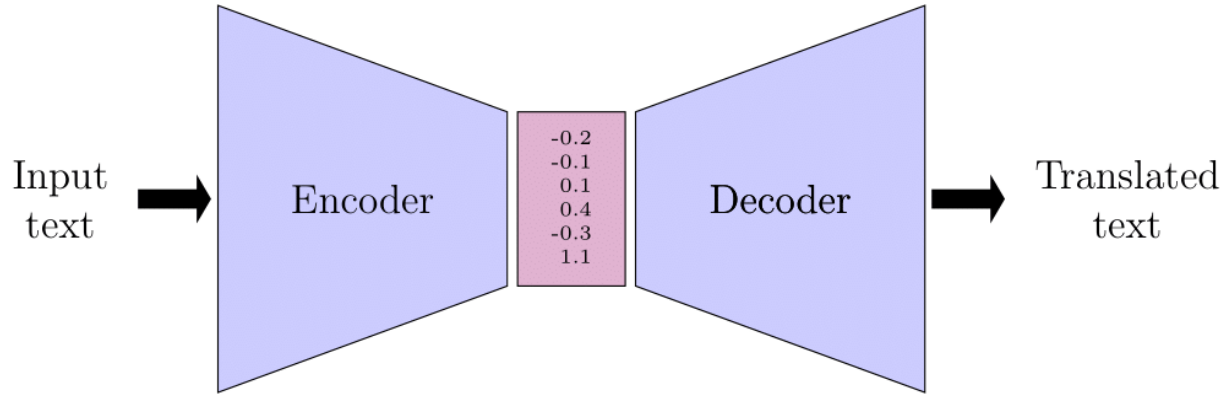

Figure 2: Diagram of encoder-decoder style NMT system

This paper is structured as follows: Section 2 provides a background on Irish MT efforts to date. Section 3 discusses related work in the MT field, while Section 4 discusses features of the Irish language which can pose a challenge for MT. Section 5 describes the collection of monolingual and bilingual datasets to be used in MT experiments. In Section 6 we give a description of the backtranslation methodology used, and provide results of MT experiments involving backtranslation in Section 7. Finally, in Section 8 we offer some conclusions and provide an insight into possible areas of future work.

2. Background

As explained in Section 1, the availability of language data (monolingual and bilingual text) is of huge importance in building MT systems. Despite language collection efforts, EN-GA MT still suffers from a lack of data. This is sometimes described as *data sparsity*. Data sparsity is a two-fold problem for EN-GA MT. The first part pertains to the availability of language data. As mentioned earlier, both SMT and NMT rely on large amounts of parallel data. If the amount of data available to train the translation models with is sparse, then it is much more difficult to achieve quality translations. The other aspect of data sparsity that affects EN-GA MT relates to the fact that Irish is more heavily inflected than English, i.e. for one Irish lemma there could be

many surface forms. This can lead to a 'one-to-many' translation situation, whereby a single word in English could have many different Irish translations, depending on the context. This makes it more difficult for MT systems to 'learn' the correct translations, further highlighting the need for more EN-GA data.

One method commonly used to improve the quality of low-resource MT is to focus on tailoring a system to a particular domain.[2] A recent pilot study that involved the development of the *Tapadóir*[3] system has shown domain-tailored MT to be useful in the post-editing environment of an official Irish government department,[4] where the translation of EN-GA documents has been facilitated by SMT (see Dowling et al. 2016 for a detailed description). The success of this domain-specific SMT system is due in part to the availability of high-quality parallel data in this particular domain (Irish public administration). This exploratory study has shown that introducing MT into the workflow of English-Irish translators within an Irish government department is beneficial, yet contains scope for improvement.

Another method for dealing with data sparsity is making concerted efforts with respect to data collection. Within a European setting, the availability of bilingual EN-GA text is limited. There is a derogation currently in place with respect to the production of Irish language content within the EU, meaning that very little Irish language content is produced in comparison to other official EU languages (Interinstitutional style guide 2011). This derogation is due to be lifted at the end of 2021, at which point there will be a significant increase in the number of translators needed, compared to current requirements. While efforts are underway to increase the number of translators available (e.g. a new MA in translation[5]) meeting the quota necessary to produce this volume of translations will still prove to be challenging without adequate technology. The NMT-based eTranslation system in use by the Directorate General for Translation (DGT) is still in its

2 Typical domains include legal, news, literary, etc.

3 From the Irish '*tapa*' meaning 'fast' and the nominal suffix 'óir'.

4 The Department of Culture, Heritage and the Gaeltacht (DCHG). DCHG is the Irish government department responsible for promoting, protecting and advancing the use of the Irish language.

5 http://www.nuigalway.ie/courses/taught-postgraduate-courses/ translation-studies.html

early days of uptake, partly due to the lack of data available to train it (CEF Digital 2019).

Unable to rely on using official EU translation data alone to train EN-GA MT systems, other methods must be employed in order to develop the size of GA datasets. Ireland's participation in the European Language Resource Coordination (ELRC)[6] is facilitating this development. The ELRC is a European Commission-led effort to collect language resources for official EU languages, with a view to ensuring that all EU Digital Service infrastructures (such as eJustice, eProcurement etc) will be accessible in all EU languages via the eTranslation system. The ELRC has also made the collected data available within Ireland to improve MT at a national level. The extent to which the ELRC data collection efforts impact our research is discussed in Section 5.2.

3. Related Work in MT

As discussed in Section 1, currently the primary focus of the application of Irish MT is within the context of a professional translation workflow (involving post-editing by human translators), and as such, progress in this area in terms of advances in state-of-the-art approaches is of interest to us. For many years, there have been extensive studies to show how the integration of MT within such a workflow (often complementary to the use of translation memory (TM) tools) improves productivity, both in industry-based and in academic-based research (e.g. Arenas 2008; Etchegoyhen et al. 2014).

With the introduction of deep learning methods in recent years, we have witnessed a breakthrough in the field of MT. NMT is increasingly showing more positive results and has been seen to outperform SMT across a number of language pairs and domains (Bojar et al. 2016). This has led to the need for subsequent studies examining the differences between the impact that SMT and NMT have within such a setting. For example, Bentivogli et al. (2016) carried out a small-scale study on post-editing of English-German translated TED talks, and concluded that NMT had led to significant improvements in the translation quality. Bojar et al. (2016) report a significant step forward using NMT instead of SMT in automatic post-editing tasks (in terms of lexical

6 http://www.lr-coordination.eu/

selection, word order, etc.) at the Conference on Statistical Machine Translation (WMT16).[7] More recently, Castilho et al. (2017) carried out a more extensive quantitative and qualitative comparative evaluation of SMT and NMT using automatic metrics and professional translators. Results were mixed overall. They varied from showing positive results for NMT in terms of improved (perceived) fluency and errors, to achieving no particular gains over SMT at document level for post-editing. While these studies were carried out on better resourced language pairs (English-German, -Portuguese, -Russian and -Greek), they are still highly relevant in indicating the potential impact that the change in MT approaches can have in real-life translation scenarios.

Despite promising results, the positive impact of NMT is not being felt across the board. In previous work (Dowling et al. 2018) we began to explore whether NMT is a viable paradigm for EN-GA MT and concluded that while there is great potential for EN-GA NMT, SMT currently outperforms it in a domain-specific scenario. As Koehn and Knowles (2017) highlight, current NMT systems can face a number of challenges when dealing with specific tasks. These challenges include low-resource languages, low-frequency words arising from inflection, long sentences, and out-of-domain texts. Sennrich et al. (2016) take steps toward addressing the challenge of MT in a low-resource scenario. They present the use of backtranslation to create artificial bilingual corpora with which to train MT systems. Poncelas et al. (2018) further this strand of research by investigating the effects of this type of data on NMT systems. In this article we investigate the impact of backtranslation on EN-GA MT and identify whether it is a viable method for artificial data creation.

4. Features of the Irish language that can pose a challenge for MT

As well as being a low-resource language, there are some linguistic features of the Irish language that pose a challenge for MT. In this section we outline some of the linguistic-based challenges that we have attempted to address.

One feature of Irish that can have an effect on EN-GA MT quality is its inflected nature. Irish words

7 http://www.statmt.org/wmt16/

can inflect for number, tense, person, case, mood and gender. Some ways that Irish words inflect include lenition (the infixing of a 'h' after the first consonant), eclipsis (a type of initial mutation where a letter is added to the beginning of the word) and slenderisation (changing a 'broad' vowel ('a', 'o' or 'u') to a 'slender' vowel ('i' or 'e')). A typical example of noun inflection can be seen in Example (1), using the feminine noun *'beach'*, meaning bee.[8] As discussed in Section 2, this can lead to data sparsity wherein inflected words are seen infrequently in the training data and the incorrect inflection is often produced in the MT output. Inflection (*b**h**each*), eclipsis (**m**beach) and slenderisation (be**i**ch**e**) can all be seen in this example.

(1)

beach	bee/a bee
*an b**h**each*	the bee
*beach**a***	bees
*dath na be**i**ch**e***	the colour of the bee
*dath na **m**beach*	the colour of the bees

Inflection has also been shown to have an impact on automatic MT evaluation metrics such as BLEU (Papineni et al. 2002). Test data, a bilingual corpus deemed to be 'gold standard' quality translation, is used to generate BLEU scores. The source language of the test data is translated, producing an MT output in the target language. This is then compared to the target-language portion of the test data to generate a BLEU score of between 0 and 100. The higher the score, the closer the MT output is to the reference translation. A shortcoming of BLEU is that it considers inflected words as being wholly different from their uninflected counterparts, and can sometimes penalise translation output too harshly as a result (Callison-Burch et al. 2006). For example, if an MT system were to output *'an beach*'* as the translation for 'the bee' and the reference contained *'an bheach*' then BLEU would consider *'beach'* as being completely incorrect. It would score it identically if the MT output contained completely unrelated words such as '*an fear*' ('the man'), *'an cáca'* ('the cake'), *'an guthán* ' ('the phone'), etc., even though it is clear to an Irish

8 For clarity, the inflection markers (letters) in each example are displayed in bold

speaker that '*an beach'* is more correct and would require much less post-editing effort. In this study we take steps towards addressing data sparsity by increasing the size of our datasets through the creation of artificial data with backtranslation.

Inflection can also be seen in Irish verbs. Example (2) shows the regular verb *ceannaigh,* 'to buy', inflecting for person and tense (in this case, the conditional mood). In this example, 'would buy' in English could be translated in 5 different ways in Irish, depending on the context. This is called a '*one-to-many translation'* where the MT system is expected to learn many possible translations for one input. This is another example of the aspect of data sparsity in EN-GA MT, where some but not all of the Irish forms may be present in the training data, as discussed in Section 2. One possible way to minimise the effects of cases like this would be to use byte pair encoding (BPE; see Gage 1994; Sennrich et al. 2016). BPE is a pre-training step, where words are broken into subword units. These subword units, which are generated statistically, are not necessarily morphemes. For example, after BPE 'furious green ideas' might look like '| fu | rio | us | green | id | ea | s |'. The premise of this method is that there is a higher chance of a subword being present in the training data rather than a full word, particularly if the language being translated is morphologically rich. Although the addition of BPE in previous EN-GA NMT experiments (Dowling et al. 2018) showed minimal improvements in BLEU score, it is expected that as language data resources increase, the improvement may become more substantial.

(2)	C***h***eann***óinn***	I would buy
	C***h***eann***ófá***	You would buy
	C***h***eann***ódh*** *sé/sí*	He/she would buy
	C***h***eann***óimis***	We would buy
	C***h***eann***ódh*** *sibh*	You (plural) would buy
	C***h***eann***óidís***	They would buy

Another challenge for building EN-GA MT systems is the divergent word order between English and Irish. Irish follows a verb-subject-object (VSO) sentence structure, differing from the subject-

verb-object (SVO) structure of English, as illustrated in Figure 3.

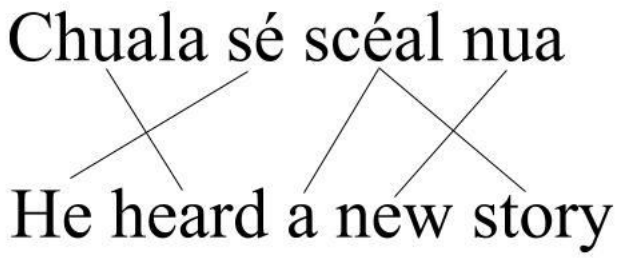

Figure 3: An example sentence highlighting the divergent word order between Irish and English

This difference in sentence structure can negatively impact MT quality, especially when translating longer sentences (Koehn and Knowles 2017). This is particularly true for SMT, which only considers a fixed number of words at a time when calculating translations. This number of words is known as an n-gram. For example, if the default 3-gram is used, the MT system will consider 3 words at a time (e.g. from Example (3) *'Chuala an fear*,' '*an fear leis,' 'fear leis an,' 'leis an bhféasóg'* ,' and so on). This can be a problem when translating between languages with divergent word orders. If an SMT system using a 3-gram model attempted to translate Example (3), *'chuala'* and 'heard' (marked in bold) would be too far away from each other to be considered as translations. In an effort to address this divergence in word order, Dowling et al. (2017) implemented a 6-gram language model in their SMT systems. We apply the same language model in our own experiments reported here.

(3)	***Chuala*** *an*	*fear*	*leis*	*an*	*bhféasóg*	*scéal*	*nua.*	
	Heard	the	man	with	the	beard	story	new.

'The man with the beard **heard** a new story.'

This issue of MT of languages with different word orders is exacerbated when translating more complex sentences. For example, in Irish, there are two forms of the verb 'to be', the substantive verb and the copula. A sentence with a copula construction can be seen in Example (4). When compared to Example (5) which uses the copula, the English sentence structures are very similar,

whereas the Irish sentences are very different, with Example (5) using the substantive verb (‘*tá’).* It is clear that this would be a difficult scenario for MT to consistently translate correctly. However, while SMT has earned the reputation of outputting sometimes overly-direct translations, NMT is much better equipped to identify context-dependent patterns. This leads us to predict that NMT may be more successful than SMT at identifying the cases in which the copula, rather than the substantive verb, is the correct structure to produce. It is our hope that, by increasing parallel data through collection efforts and backtranslation, there will be sufficient EN-GA data to build NMT systems that are better equipped to identify and produce these constructions.

(4) ***Is*** *scoláire* *í*
Is scholar she
‘She **is** a scholar.’

(5) ***Tá*** *sí* *beo*
Is she alive
‘She **is** alive.’

We note that this is not an exhaustive list of the complex Irish features that can have an impact on MT, rather the specific ones that we have taken steps to address to date.

5. Data

As mentioned in Section 2, both SMT and NMT rely on large amounts of bilingual data to train a system. A lack of suitable training data can lead to poor quality MT systems. In addition, many recent techniques for improving MT quality, especially those pertaining to NMT, require a large amount of data before their benefits can be seen. Therefore, a crucial step in the development of EN-GA MT is to gather and curate suitable EN-GA data resources.

5.1. Baseline resources

We take our previous work on the Tapadóir system as a baseline comparison for our studies (Dowling et al. 2016). During the Tapadóir SMT project, parallel and monolingual datasets were identified and gathered (see Table 1) in order to build a baseline system. A baseline system is necessary in MT experiments to monitor the effect of various datasets and changes in MT approaches on the output. We describe their datasets here to fully explain the basis of comparison and the data used in this current study.

The resources available from that project are the following: (i) the largest corpora in the Tapadóir project, are the TMs (bilingual files containing previous translations by inhouse translators) provided by DCHG. This corpus is referred to as **DCHG** in Table 1: (ii) Corpas Comhthreomhar Gaeilge-Bearla (**CCGB**) is a bilingual dataset obtained through web-crawling, and is available for download online[9]; (iii) Parallel texts from two EU bodies – the Digital Corpus of the European Parliament and Directorate General for Translation, Translation Memories – are also publicly available (these datasets are referred to collectively as **EU** in Table 1); (iv) Another dataset available is the Parallel English–Irish corpus of legal texts (referred to as **'Gaois'** in Table 1) made available online by the Department of Justice.[10] The language used is very technical and contains much 'legalese', or legal jargon. (v) As well as this, 10,000 parallel sentences were crawled from the Citizens Information website,[11] referred to as **CitizensInfo** in Table 1. All datasets apart from **DCHG** are either publicly available online or available to webcrawl. It is our hope that in the future, following consultation with the data holders and anonymisation of the data, we may be able to publish the DCHG dataset. Dowling et al. (2016) provide detailed results of building MT systems using these datasets.

Corpus	**# of words (GA)**	**# of sentences**
DCHG	440,035	29,000

9 https://github.com/kscanne/ccgb
10 https://www.gaois.ie/crp/en/data/
11 http://www.citizensinformation.ie

CCGB	113,889	6,000
EU	439,262	29,000
Gaois	1,526,498	89,000
CitizensInfo	183,999	10,000
TOTAL	**2,703,683**	**163,000**

Table 1: Baseline datasets

5.2. Tapadóir version 2 resources

While the resources mentioned in Section 5.1 can be used to build a promising baseline system, these datasets were expanded through further data-gathering efforts by Dowling et al. (2018). This was achieved through 1) directly contacting organisations which deal with Irish-language content and 2) web-crawling.

5.2.1. Through direct contact with organisations

DCHG: Following on from the existing collaboration with DCHG during the Tapadóir project, DCHG continue to provide us with TMs created by their team of in-house translators. This data, translated by professional translators within the setting of a government department, can be described as being 'gold-standard', i.e. of a high enough quality that it is suitable for use as both training and testing datasets for MT. As well as the original DCHG corpus (see Table 1) two additional corpora have been collected from DCHG. These are referred to as **DCHG†** and **DCHG††** respectively, in Table 2.

ELRC: As one of the 24 official EU languages, Irish is included in the ELRC project, which seeks to gather language technology resources in order to provide suitable digital facilities for all European citizens. With the added weight of an official EU project, the ELRC representatives in Ireland contacted Irish language organisations and public bodies that have obligations to provide Irish-language content in order to request language data from them. This involved the organisation of two workshops aimed at educating language holders on the value of language technology and resource sharing, as well as directly contacting or visiting organisations to aid

them in identifying and sharing corpora. The data they collected was extremely varied, both in terms of quantity and format. While only a small number of datasets collected during the ELRC project have been included in our MT experiments to date (data from the University Times,[12] referred to as **'UT'** in Table 2 and data from Conradh na Gaeilge,[13] referred to as **'CnaG'** in Table 2), this project, now in its second phase, is expected to continue to identify language data holders and increase the amount of monolingual and bilingual data available for EN-GA. The National Corpus of Ireland (NCI) is a large monolingual corpus of Irish which Foras na Gaeilge contributed to the ELRC project. It contains 1,994,081 Irish sentences (see Table 2). The domain of this corpus can be described as 'mixed' or 'general', as it contains poetry, literature and news articles, among many other types of content.

5.2.2. Web-crawling

Web-crawling is a common method for collecting bilingual data from websites for language pairs that may be lacking in resources. With both Irish and English as official languages of Ireland, many public websites have an obligation to provide online content bilingually. In Dowling et al. (2016), we report on a list of websites that could contain bilingual content and crawled them using the ILSP crawler (see Papavassiliou et al. 2013). The resulting corpora were reported to be often of mixed quality; common issues included misalignment (mismatched translations), comparable (similar content rather than truly parallel) content, noisy data (containing HTML markup, typos, etc.) and crawling failure. In addition, while the crawler relies on consistency in webpage labelling that clearly indicates the content's language, this is not the case for many Irish websites. A pre-processing stage was introduced before adding this data to be used for MT development. This stage involved full cleaning (removal of formatting such as XML or HTML tags) and accurate manual alignment. A further 4,028 parallel sentences from various sources were obtained through crawling (referred to as **'Crawled'**). Their additional crawled datasets are referred to as **'IT'**[14] and **'Teagasc'**[15] in Table 2.

12 University Times is the student newspaper in Trinity College Dublin.
13 Conradh an Gaeilge is an organisation which promotes the use of the Irish language.
14 The Irish Times (IT) is a national newspaper in Ireland.
15 Teagasc is the state agency providing research, advisory and education in agriculture, horticulture, food

Corpus	**# of words (GA)**	**# of sentences**
DCHG†	243,372	13,500
DCHG††	402,210	23,714
UT	15,377	598
CnaG	21,365	1,367
Crawled	70,773	4,028
Teagasc	32,908	1,473
IT	57,314	2,694
EU	483,149	29,445
NCI	18,964,885	1,994,081
TOTAL	**17,861,353**	**2,070,900**

Table 2: Size of additional resources gathered

It can be seen from the final row in Table 2 a total of 17,861,353 GA words of data were used for MT training. Results from building engines using combinations of this data can be seen in Dowling et al. (2016), (2017) and (2018).

5.3. Test data

In order to test the Tapadóir system, in previous work we held out a random sample of 1,500 sentence pairs received from DCHG from the training set to form the test set. The test set is therefore domain-specific, and representative of the type of texts the Tapadóir system is expected to translate (letters, reports, press releases, etc.). We use the same test set in our current study in order to make a clear comparison.

6. Back Translation Experiment Set-Up

While data collection efforts are extremely important in the context of resource-poor MT, it can

and rural development in Ireland.

be time-consuming and not always result in an improvement in the quality of MT output. The creation of artificial data is a quick, experimental way of increasing the amount of bilingual data available for a language pair. This section describes the methodology used for the creation of EN-GA artificial data through backtranslation.

6.1. Data

The data used for building the GA-EN SMT system is listed in Table 2 (the same as that used by Dowling et al. (2018), which is in the public administration domain, specifically DCHG). For consistency, the same datasets were used in the creation of the baseline NMT system, with the exception of monolingual data. Monolingual data is not usually used in the training of an NMT system. However, it is required for use in backtranslation experiments. It is preferable to use as large a corpus as possible in these experiments to maximise the amount of artificial parallel data created.

6.2. Setup and Methodology

Figure 4 illustrates a simplified version of the backtranslation method implemented. In Step (1), authentic parallel data (bilingual text usually translated by professional translators) is used to train a GA-EN SMT system. As shown in Step (2), a large monolingual corpus is then machine-translated using this SMT system. This creates an artificial parallel corpus: genuine Irish language text on one side and machine-translated English text on the other. Finally, in Step (3), this artificial corpus is used to train an EN-GA NMT system. Being the largest corpus of monolingual Irish data, the NCI corpus (see Table 2) was identified as a suitable starting point for the creation of artificial bilingual data. Previous experiments (Dowling et al. 2018) suggest that SMT outperforms NMT when dealing with this language pair. For this reason, we translate the NCI corpus using a GA-EN SMT system. We then train EN-GA NMT systems using differing ratios of artificial data to authentic data. Applying a similar method to that used by Poncelas et al. (2018), we first begin with a 1:1 ratio of artificial versus authentic training data, and iteratively add more data until the entire monolingual corpus has been fully translated and all artificial parallel data has been added to the training data.

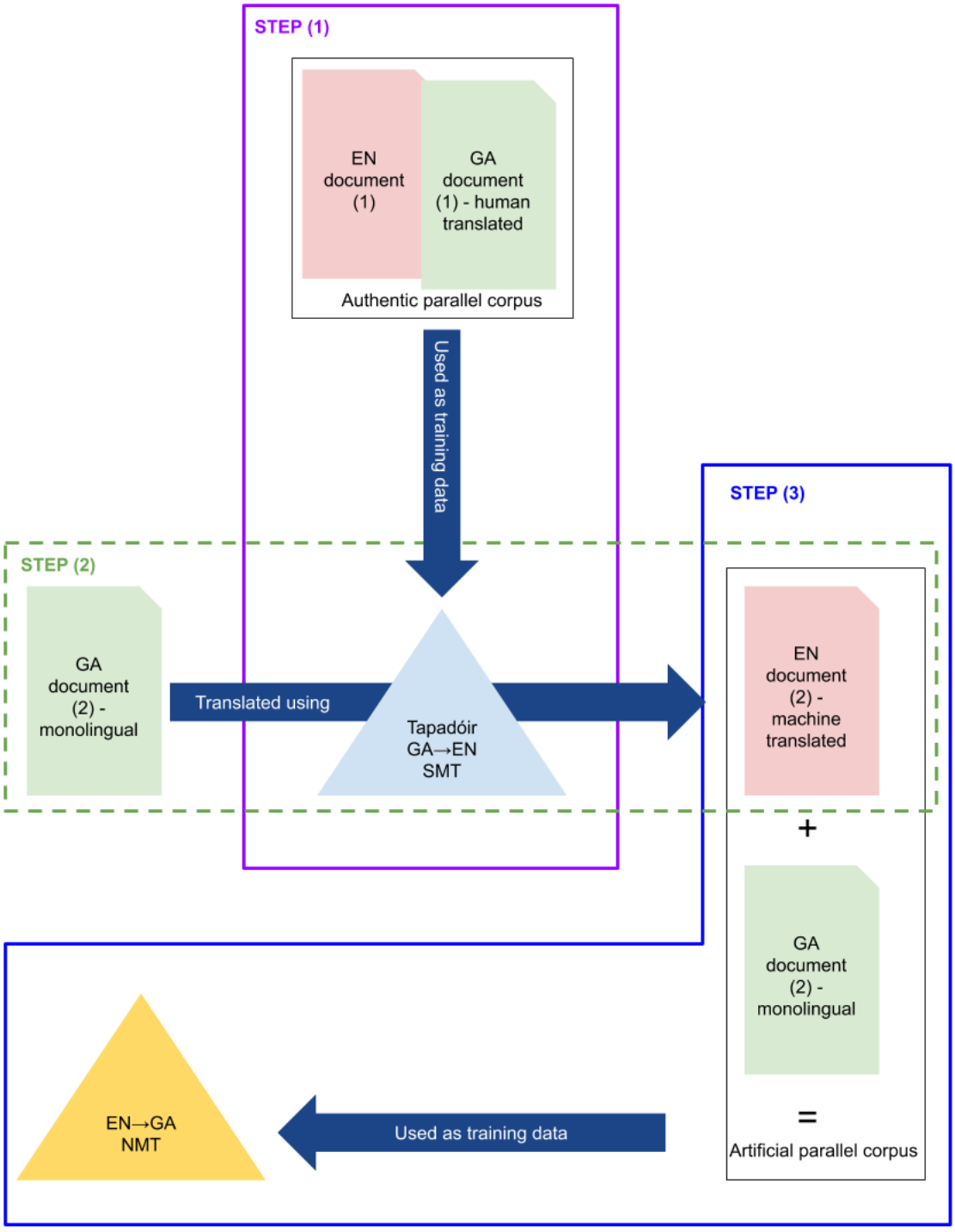

Figure 4: Simplified backtranslation diagram

7. Results and Preliminary Analysis

The test set described in Section 5.3 was used to obtain BLEU scores for each backtranslation experiment. The results of these experiments are shown in Figure 5, in which we also provide the baseline NMT score with no artificial data added (0:1 ratio) as reported by Dowling et al. (2018). These results show that, contrary to related research, the inclusion of back-translated data does not improve the BLEU score of EN-GA NMT when using these datasets and configuration. It can be seen from both Table 3 and Figure 5, that the higher the ratio of artificial to authentic data, the more the MT output decreases in BLEU score. As a marker of sufficient BLEU quality, Escartin et al. (2015) indicate that for the Spanish-English pair, a BLEU score of 45+ can increase translator productivity. Although these experiments have not been repeated with EN-GA, we can take this score as a rough guideline. The BLEU scores achieved using backtranslation fall below this threshold, and continue to fall as more artificial data is added. This could indicate that the MT systems trained using backtranslation data would not be suitable in the post-editing workflow of a professional translator.

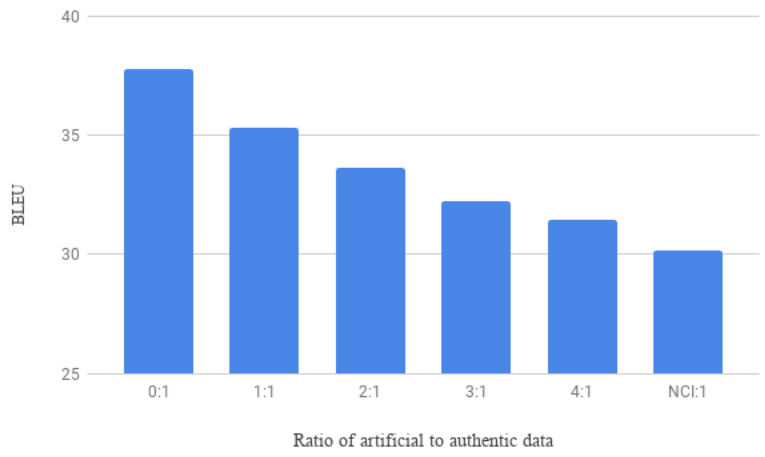

Figure 5: Barchart displaying BLEU scores of backtranslation MT systems, with the NMT system from Dowling et al., 2018 as a comparison

MT SYSTEM	**BLEU**
SMT-baseline	39.36
SMT-best	46.44
NMT	37.77
BT 1:1	35.29
BT 2:1	33.61
BT 3:1	32.24
BT 4:1	31.46
BT NCI:1	30.18

Table 3: BLEU scores for backtranslation experiments as well as earlier SMT and NMT systems performed by Dowling et al. 2016, 2017 and 2018 for comparison

7.1. Sentence-level BLEU analysis

In order to gain a preliminary insight into specific changes in the MT output brought about by the introduction of backtranslated data, we performed a sentence-level BLEU analysis. This means that, rather than solely generating an overall BLEU score for the test document (as is the norm), an individual BLEU score is given for each sentence. This information can then be used to identify sentences with the biggest discrepancy in BLEU scores. In Example (6),[16] we see the evolution of a machine-translated sentence as more artificial data is introduced to the NMT training phase. It can be seen that the baseline system with no artificial data (0:1) matches the reference exactly, and so achieves a perfect BLEU score. With the first introduction of artificial data (1:1) we see that the translation output changes for the worse ('i', 'in' instead of 'mar', 'because'). This leads to a more literal translation, which is interesting because in general it is reported that NMT is better equipped to produce fluent, rather than literal, translations (Castilho et al. 2017). The next highest ratio of artificial data (2:1) shows a similar output, though slightly more grammatical ('i' is inflected to be 'in' before a vowel). Ratios 3:1, 4:1 and NCI:1 (just over 5:1) see the semantics of the sentence completely changed (the NCI:1 output could be roughly translated as 'a summary

16 For all examples, the ⇓ and ⇑ symbols indicate a drop or increase respectively in BLEU score over the authentic data (0:1) BLEU score

cannot be given for these circumstances'). These examples highlight a common issue in NMT - the output looks perfectly fluent but actually displays a completely different meaning to the source text (Castilho et al. 2017).

(6)

Source: in summary, the Department's purpose is:

Irish reference: *mar achoimre, is é cuspóir na Roinne:*

0:1 (authentic data only): *mar achoimre, is é cuspóir na Roinne:*	BLEU = 100
1:1: *i achoimre, is é cuspóir na Roinne:*	BLEU ⇓ 10
2:1: *in achoimre, is é cuspóir na Roinne:*	BLEU ⇓ 10
3:1: *ní feidir achoimre a dhéanamh ar an méid sin*	BLEU ⇓ 90.06
4:1: *ní mor achoimre a thabhairt ar an gceist seo*	BLEU ⇓ 90.06
NCI:1: *ní feidir achoimre a thabhairt ar na cúinsí seo*	BLEU ⇓ 70.12

Despite somewhat negative results, in Example (7) we see an occasion where backtranslation has improved the NMT output, both in terms of automatic evaluation and human analysis. With authentic data only, the NMT system incorrectly translates 'description' as *tuairisc*, 'report.' The first addition of artificial data (1:1) produces the MT output *cur sío*s which is an exact match of the human-translated reference. This is echoed with in the next addition of artificial data (2:1), but changes to *cur síos ar*, 'description of' in experiments with ratios 3:1 and 4:1. This is another example of NMT appearing fluent (*ar* is the appropriate preposition in this situation) but containing differing semantics to the source. However, in the final addition of data (NCI:1), the MT again outputs the correct translation. This raises the question 'how much artificial data harms the MT output, and how much benefits it?' This could be an indication that if a greater amount of artificial data were added a higher level of MT accuracy could be gained.

(7)

Source: description.

Irish reference: *cur síos.*

0:1 (authentic data only): tuairisc.	BLEU = 30.33
1:1: *cur síos .*	BLEU ⇑ 69.67
2:1: *cur síos .*	BLEU ⇑ 69.67
3:1: *cur síos ar .*	BLEU ⇑ 90
4:1: *cur síos ar .*	BLEU ⇑ 36.34
NCI:1: *cur síos .*	BLEU ⇑ 69.67

8. Conclusions and Future Work

In this paper, we have presented preliminary results of the use of backtranslation as a means of generating artificial data for EN-GA MT. Despite previous studies on other language pairs reporting the contrary, we have shown from both automatic and preliminary linguistic evaluation of the MT output that backtranslation was not successful in improving EN-GA MT using the current configuration. We can hypothesise a number of reasons for this. Firstly, perhaps our synthetic datasets were too out-of-domain, given that the NCI corpus contains a mixture of domains (i.e. literature, legal, news, etc.) and may differ too much from our domain-specific test set. Possible future work to address this issue could be to identify a monolingual dataset that is closer in domain to text from DCHG and rerun the experiments using that as a basis for the artificial parallel corpus. This could provide further insights into the importance of data selection and domain in MT.

Secondly, the original training dataset available is much smaller than those used by Poncelas et al. (2018) (1m sentences). To this end, the most obvious approach is to continue to experiment as the ELRC collects more data. The recent launch of the related project, the European Language Resource Initiative,[17] will further encourage language resource sharing from Irish public bodies which will result in a larger training corpus.

Thirdly, there may in fact be improvements in quality, but the BLEU evaluation metric is clearly not equipped to identify them. It could be seen from Example (7) that it is in fact possible for

[17] https://elri.dcu.ie/ga-ie/

backtranslation to improve some parts of the MT output. An empirical study by Shterionov et al. (2018) shows that the disconnect between BLEU and human evaluation may be as much as 50%. Way et al. (2019, to appear) highlight the shortcomings of BLEU and conjecture that other methods of evaluation will, particularly those tuned to NMT, will be necessary in the future. With BLEU offering only a small insight into the quality of MT, it will be important in future work to experiment with other automatic metrics, for example ChrF3 (Popovic 2015) a character-based metric that penalises less harshly than BLEU for small inflection errors. As well as automatic metrics, it will be vital to use more human evaluation to gain insights into the MT output quality. Human evaluation can be used to ensure that the MT systems designed for public administration use will be optimised to enhance the task of specific human translators, and will not merely be tuned to automatic metrics. There are a variety of methods for human evaluation, not merely in the assessment of the quality of the output, but also in terms of assessing the suitability of MT systems in a post-editing environment (through the use of eye-tracking technology, calculating edit-distance, key-strokes etc., e.g. Castilho and Guerberof 2018).

In terms of demand for EN-GA MT, it is important to note that the derogation on the production of Irish-language documents within the EU is due to lift in 2021. Both nationally and at a European level, those tasked with EN-GA translation will need to look to technology to help increase productivity. It is vital, therefore, that MT resources are well-developed, up-to-date and designed accordingly to meet this demand.

Acknowledgments

This research was conducted at the ADAPT SFI Research Centre at Dublin City University. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant # 13/RC/2106. This work was part-funded by the Department of Culture, Heritage and the Gaeltacht. We thank the reviewers for their valuable feedback and advice.

References

Arenas, A. G. 2008. Productivity and quality in the post-editing of outputs from translation memories and machine translation. *Localisation Focus,* 7(1), pp. 11–21.

Bahdanau, D., K. Cho and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of International Conference on Learning Representations.* San Diego, California, USA. 15 pages.

Bentivogli, L., A. Bisazza, M. Cettolo, and M. Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 257–267. Austin, Texas: Association for Computational Linguistics.

Bojar, O., R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Neveol, M. Neves, M. Popel, M. Post, R. Rubino, C. Scarton, L. Specia, M. Turchi, K. Verspoor, and M. Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers,* pp. 131–198. Berlin: Association for Computational Linguistics.

Callison-Burch, C., M. Osborne, and P. Koehn. 2006. Re-evaluation the role of BLEU in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics,* pp. 249–256. Trento, Italy: Association for Computational Linguistics.

Castilho, S., J. Moorkens, F. Gaspari, R. Sennrich, V. Sosoni, Y. Georgakopoulou, P. Lohar, A. Way, A. V. Miceli Barone, and M. Gialama. 2017. A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In *Proceedings of Machine Translation Summit XVI*, pp. 116–131. Nagoya, Japan: Association for Computational Linguistics.

CEF Digital. 2019. What is eTranslation. [online] Available at: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/What+is+eTranslation.

Castilho, S. and A. Guerberof Arenas. 2018. Reading comprehension of machine translation output: what makes for a better read? In *21st Annual Conference of the European*

Association for Machine Translation. pp 79 -88. Alicante, Spain: Association for Computational Linguistics.

Dowling, M., L. Cassidy, E. Maguire, T. Lynn, A. Srivastava, and J. Judge. 2015. Tapadóir: Developing a statistical machine translation engine and associated resources for Irish. In *Proceedings of the Fourth LRL Workshop: "Language Technologies in support of Less-Resourced Languages"*, pages 314–318. Poznan, Poland: Fundacja Uniwersytetu Im. Adama Mickiewicza.

Dowling, M., T. Lynn, Y. Graham, and J. Judge. 2016. English to Irish machine translation with automatic post-editing. In *Proceedings of the Second Celtic Language Technology Workshop*, pp. 42–54. Paris: Association for Computational Linguistics.

Dowling, M., T. Lynn, A. Poncelas, and A. Way. 2018. SMT versus NMT: Preliminary comparisons for Irish. In *Technologies for MT of Low Resource Languages (LoResMT 2018), AMTA*, pp. 12–20. Boston: Association for Computational Linguistics.

Etchegoyhen, T., L. Bywood, M. Fishel, P. Georgakopoulou, J. Jiang, G. van Loenhout, A. del Pozo, M. S. Maucec, A. Turner, and M. Volk. 2014. Machine translation for subtitling: a large scale evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation,* pp. 46–53. Reykjavik: European Language Resources Association (ELRA).

Forcada, M. L. 2017. Making sense of neural machine translation. *Translation spaces*, 6(2), pp. 291–309.

Gage, P. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2), pp. 23–38.

Hearne, M., & A. Way. 2011. Statistical machine translation: a guide for linguists and translators. *Language and Linguistics Compass*, 5(5), pp. 205–226.

Interinstitutional style guide. (2011). Brussels: Publications Office of the European Union, p.116.

Koehn, P. and R. Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pp. 28–39. Berlin: Association for Computational Linguistics.

Neubig, G. 2017. *Neural machine translation and sequence-to-sequence models: A tutorial.* arXiv preprint arXiv:1703.01619. Accessible at https://arxiv.org/abs/1703.01619.

Papavassiliou, V., P. Prokopidis, and G. Thurmair. 2013. A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pp. 43–51, Sofia, Bulgaria: Association for Computational Linguistics.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU:: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Philadelphia, Pennsylvania: Association for Computational Linguistics.

Poncelas, A., D. Shterionov, A. Way, G. M. de Buy Wenniger, and P. Passban. 2018. Investigating backtranslation in neural machine translation. In *21st Annual Conference of the European Association for Machine Translation*, pp. 249–258. Alicante, Spain: Association for Computational Linguistics.

Popović, M. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395. Lisbon: Association for Computational Linguistics.

Sennrich, R., B. Haddow, and A. Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1: Long Papers, pp. 86–96. Berlin: Association for Computational Linguistics..

Sennrich, R., B. Haddow, and A. Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1: Long Papers, pp. 1715–1725. Berlin: Association for Computational Linguistics.

Shterionov, D., R. Superbo, P. Nagle, L. Casanellas, T. O'Dowd, and A. Way. 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation*, 32(3), pp. 217–235.

Sutskever, I., O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of Neural Information Processing Systems 2014*, pp. 3104–3112. Montréal: Advances in Neural Information Processing Systems 27.